

# Some Notes on Sample Selection Models

Victor Aguirregabiria

Department of Economics. University of Toronto

20. May 2009

# Some Notes on Sample Selection Models

Victor Aguirregabiria
University of Toronto

June 27, 2009

**Abstract**

Sample selection problems are pervasive when working with micro economic models and datasets of individuals, households or firms. During the last three decades, there have been very significant developments in this area of econometrics. Different type of models have been proposed and used in empirical applications. And new estimation and inference methods, both parametric and semiparametric, have been developed. These notes provide a brief introduction to this large literature.

**Keywords:** Sample selection. Censored regression model. Truncated regression model. Treatment effects. Semiparametric methods.
**JEL codes:** C10, C35, C63.

**Corresponding Author:** Victor Aguirregabiria. Address: 150 St. George Street. Toronto, ON, M5S 3G7. Phone: (416) 978-4358. E-mail: victor.aguirregabiria@utoronto.ca

# 1 Introduction

Consider the regression model,

$$Y^* = X^*\beta + \varepsilon \tag{1}$$

where $Y^*$ and $\varepsilon$ are scalar random variables, $X^*$ is a $1 \times K$ vector of random variables, and $\beta$ is a $K \times 1$ vector of parameters. The error term $\varepsilon$ is mean independent of $X^*$, and the matrix $E(X^{*\prime}X^*)$ is full rank. Therefore, given a random sample of the variables $\{Y^*, X^*\}$, the OLS estimator is consistent and asymptotically normal. The key feature of sample selection models is that the researcher does not observe a random sample of the variables $\{Y^*, X^*\}$. Instead, the researcher observes a random sample of variables $\{Y, X\}$ which are related to but they are different to $\{Y^*, X^*\}$. The variables $\{Y^*, X^*\}$ are called latent variables. Given

a random sample of $\{Y, X\}$, we are interested in the consistent estimation of $\beta$.[1] We have different classes of sample selection models depending on the relationship between the latent variables and the observed variables

*(a) Truncated Regression Model.* Let $c$ be a known constant. If $Y$ is truncated to the left at $c$:

$$(Y, X) = \{ \ (Y^*, X^*) \mid Y^* > c \ \} \tag{2}$$

If $Y$ is truncated to the right at $c$:

$$(Y, X) = \{ \ (Y^*, X^*) \mid Y^* > c \ \} \tag{3}$$

Then, a random sample of $(Y, X)$ is not random sample neither of $Y^*$ nor of $X^*$.[2]

**Example 1:** Consider the log-wage equation, $W^* = X^*\beta + \varepsilon$, where $W^*$ is the logarithm of an individual's wage, and $X^*$ is a vector of observed human capital characteristics. Suppose that, for reasons of confidentiality, our data set does not report any information (neither of wages nor of individual characteristics) for individuals with an hourly wage greater than $\$800/hour$. Therefore, we observe the variables $(W, X)$ such that $(W, X) = \{(W^*, X^*) \mid W^* < \ln(800)\}$. In this case, we say that the dependent variable is truncated to the right, and we have a truncated regression model because neither $W^*$ nor $X^*$ are observed when the wage is greater than $\$800/hour$.

Let $f_{Y^*}$ and $F_{Y^*}$ be the density function (PDF) and the cumulative distribution function (CDF) of $Y^*$, respectively. If $Y$ is left truncated at $c$, then the PDF of $Y$ is,

$$f_Y(y) = \begin{cases} 0 & \text{if} \quad y \leq c \\ \dfrac{f_{Y^*}(y)}{1 - F_{Y^*}(c)} & \text{if} \quad y > c \end{cases} \tag{4}$$

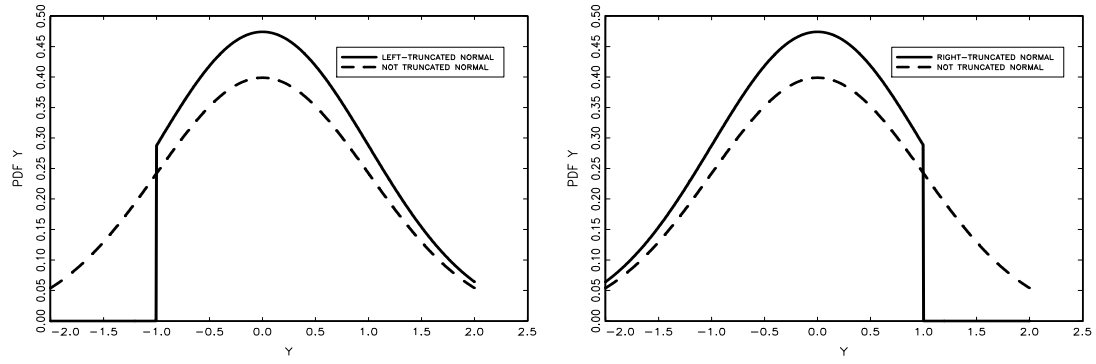
[1] In some applications, we may be also interested in the estimation of the distribution function of the error term $\varepsilon$.

[2] Only if $Y^*$ and $X^*$ are independently distributed, then a random sample of $(Y, X)$ implies a random sample of $X^*$.

If $Y$ is right truncated at $c$, then

$$f_Y(y) = \begin{cases} \dfrac{f_{Y^*}(y)}{F_{Y^*}(c)} & \text{if} \quad y < c \\ \\ 0 & \text{if} \quad y \geq c \end{cases} \tag{5}$$

The following figures present the density functions of left truncated and right truncated normal random variables.




*(b) Censored Regression Model (or Tobit Model).* The main difference between this model and the truncated regression model is that now we have a random sample of the exogenous regressors $X^*$. That is, the random variables $X$ and $X^*$ are identical. For the dependent variable, if $Y$ is left censored we have that:

$$Y = \max\left[Y^* \ ; \ c\right] = \begin{cases} c & \text{if} \quad Y^* \leq c \\ \\ Y^* & \text{if} \quad Y^* > c \end{cases} \tag{6}$$

If $Y$ is right censored, then

$$Y = \min\left[Y^* \ ; \ c\right] = \begin{cases} Y^* & \text{if} \quad Y^* < c \\ \\ c & \text{if} \quad Y^* \geq c \end{cases} \tag{7}$$

**Example 2:** Consider the log-wage equation in Example 1. Now, we have a different dataset. The data include information for every individual regardless her income level. There is a random sample of individuals with information on the $X$ variables. However, for confidentiality reasons, data on wages is top-coded. If an individual has an hourly wage lower than

$\$800/hour$, we observe the actual wage. But we do not observe wages of individuals earning more than $\$800/hour$. Therefore, for every individual in the sample we observe the censored or top-coded log-wage $W = \min[W^* \ ; \ln(800)]$. The dependent variable is censored to the right, and we have a censored regression model.

**Example 3:** Consider the following model of firm investment in a particular type of capital equipment, e.g., computers. Let $Q^*$ represent the "desired" investment of a firm according to some economic model of firm investment behavior, e.g., the amount of investment that maximizes profit if we do not restrict $Q^*$ to be positive: i.e., $Q^* = \arg\max_q \Pi(q)$, where $\Pi(q)$ is the (intertemporal) profit function. Suppose that this model implies the following regression-like equation: $Q^* = X \ \beta + \varepsilon$. The vector $X$ includes characteristics of the firm and the capital market where the firm operates such as its capital stock of the equipment, and the price of new capital. $\beta$ is a vector of parameters with clear economic interpretation within the model. We have a random sample of firms for which we observe $X$ and the amount of investment $Q$. Looking at the empirical distribution of investment $Q$, we realize that this variable is always positive and there is mass of probability at zero. These features in the distribution of investment cannot be explained by the previous regression model, unless we make very unreasonable assumptions on the distribution of $\varepsilon$. Furthermore, our model for investment assumes that $Q^*$ can be either positive or negative, and this is in contradiction with our observation of $Q$. Then, we consider the following model for $Q$, $Q = \arg\max_q \Pi(q)$ subject to $q \geq 0$. If the profit function $\Pi(q)$ is strictly concave, then it is simple to show that $Q = Q^*$ if $Q^* > 0$, and $Q = 0$ if $Q^* \leq 0$. That is, $Q = \max[Q^* \ ; \ 0]$, with $Q^* = X \ \beta + \varepsilon$. From an economic point of view this model can be interpreted as a model of irreversible investment. From an econometric point of view this is a censored regression model.

Examples 2 and 3 present two different censored regression models. It is interesting to point out some relevant differences between these two examples. They are based on very different economic and statistical assumptions. In Example 2, censoring is the result of the sampling features of our data set. The wage of individuals with wages greater than

$\$800/hour$ is not a theoretical concept, it is something that actually exists, though we do not observe it in our sample. In Example 3, censoring is a modelling assumption. Given certain features in the distribution of investment we consider that a censored regression model can be a reasonable model for this variable. The variable $Q^*$ is a theoretical concept, and we can never get a random sample of $Q^*$. However, the parameters $\beta$ can have a clear economic interpretation in this model, and they are our parameters of interest.

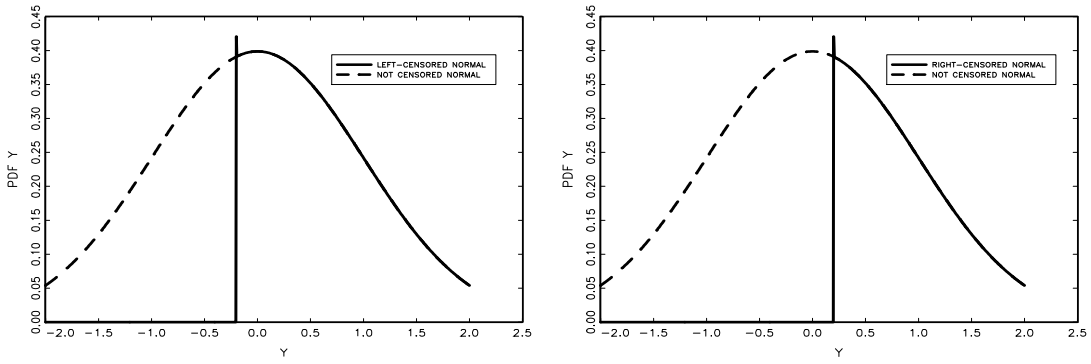
Let $f_{Y^*}$ and $F_{Y^*}$ be the PDF and the CDF of $Y^*$, respectively. If $Y$ is left censored at $c$, then the PDF of $Y$ is,

$$f_Y(y) = \begin{cases} 0 & \text{if} \quad y < c \\ F_{Y^*}(c) & \text{if} \quad y = c \\ f_{Y^*}(y) & \text{if} \quad y > c \end{cases} \tag{8}$$

If $y$ is right censored at $c$, then

$$f_Y(y) = \begin{cases} f_{Y^*}(y) & \text{if} \quad y < c \\ 1 - F_{Y^*}(c) & \text{if} \quad y = c \\ 0 & \text{if} \quad y > c \end{cases} \tag{9}$$

The following figures present the density functions of left censored and right censored normal random variables.




*(c) Sample Selection Model.* In a sample selection model, we observe $Y^*$ only for those individuals in the sample for which a certain binary variable, $D$, is equal to one. This binary

variable is not independent of $Y^*$.

$$Y = \{\ Y^* \mid D = 1\ \} \tag{10}$$

and $pdf(Y^*|D=1) \neq pdf(Y^*|D=0)$. Note that if $D$ and $Y^*$ are independent, then the random variables $Y$ and $Y^*$ are the same and there is not a sample selection problem. There are two types of sample selection models: the truncated type, and the censored type. In the truncated type, $X^*$ is also unobserved when $D = 0$. Then, in a truncated selection model:

$$(Y, X) = \{\ (Y^*, X^*) \mid D = 1\ \} \tag{11}$$

In the censored type, we have a random sample of $X^*$ (i.e., $X = X^*$). Then,

$$(Y, X) = \{\ ([Y^*|D=1]\ ,\ X^*\} \tag{12}$$

In this censored-type selection model, sometimes it is convenient to define $Y$ as follows: $Y = Y^*$ when $D = 1$, and $Y = 0$ when $D = 0$, . Or in a more compact form, $Y = DY^*$. Note that truncated and censored regression models are particular cases of the selection model. When $D = I(Y^* > c)$, the sample selection model becomes the left truncated/censored regression model, and similarly, when $D = I(Y^* < c)$, we have the right truncated/censored model.

**Example 4:** Consider again the log-wage equation in Examples 1 and 2. However, now we are not only interested in the population of individuals working but in the whole population of individuals in the labor force either working or not. Now, we interpret $W^*$ as the latent market wage of individual, and this wage exists regardless the individual is working or not. We have a random sample of individuals, working or not. Therefore, we have a random sample of $X^*$, i.e., censored-type of selection model. But we observe the market wage $W^*$ only for those individuals who are actually working. Let $D$ be the indicator of the event "the individual is working". Therefore, we have a random sample if the variable $W$, where $W = \{W^*|D = 1\}$. The working indicator $D$ depends on different factors, including human

capital characteristics observed and unobserved to the econometrician. Therefore, $D$ and $W^*$ are not independent, and we have a sample selection problem.

The specification of a sample selection model should include some assumptions on the joint distribution of $Y^*$ and $D$. A common specification is,

$$D = 1\{\ Z\ \gamma - u > 0\} \tag{13}$$

where $1\{.\}$ is the indicator function; $Z$ is a vector of observable variables; $\gamma$ is a vector of parameters; and $u$ is unobservable. The variables $(X, Z)$ are exogenous in the sense that they are independent of the disturbances $(u, \varepsilon)$. Conditional on $(X, Z)$ the unobservables $u$ and $\varepsilon$ are not independently distributed.

*(d) Generalized Sample Selection Model.* Consider the following system of $J$ linear equations,

$$\begin{array}{rcl} Y_1^* & = & X^*\ \beta_1 + \varepsilon_1 \\ Y_2^* & = & X^*\ \beta_2 + \varepsilon_2 \\ \vdots & & \vdots \\ Y_J^* & = & X^*\ \beta_J + \varepsilon_J \end{array} \tag{14}$$

Suppose that we observed a random sample of $X^*$, i.e., censored-type of sample selection model with $X = X^*$.[3] However, we do not observe all the $J$ dependent variables $\{Y_1^*, Y_2^*, ..., Y_J^*\}$ for every individual in the sample. Instead, for every individual, we observe a discrete variable $D \in \{1, 2, ..., J\}$ and a dependent variable $Y$ such that:

$$Y = \textstyle\sum_{j=1}^{J} 1(D = j)\ Y_j^* \tag{15}$$

Each individual is observe in one and only one *regime*. Importantly, the discrete variable $D$ is not independently distributed of the disturbances $\varepsilon_j' s$ in the system of linear equations.

**Example 5 (Roy Model[4]):** Consider an individual choosing between two possible occupations, 1 and 2. Suppose that this individual chooses the occupation that provides her the highest (lifetime) earnings. Given individual observable and unobservable characteristics,

[3] We can also consider a version of this model where $X^*$ is trucated for some regime $j \in \{1, 2, ..., J\}$.

[4] See Roy (1951), and Heckman and Honore (1990).

earnings in the two occupations are:

$$\begin{aligned} W_1^* &= X\ \beta_1 + \varepsilon_1 \\ W_2^* &= X\ \beta_2 + \varepsilon_2 \end{aligned} \tag{16}$$

The vector $X$ contains observable human capital characteristics such as education and labor market experience. The vectors of parameters $\beta_1$ and $\beta_2$ represent the returns to human capital characteristics in occupation 1 and 2, respectively. $\varepsilon_1$ and $\varepsilon_2$ represent returns to unobservable (for the econometrician, but not for the individual) human capital characteristics. Each individual is observed in only one occupation. Let $D$ be the indicator of the event "the individual chooses occupation 1". Therefore, the observed earnings of an individual, $W$, can be represented as:

$$W = D\ W_1^* + (1 - D)\ W_2^* \tag{17}$$

Under the assumption that individuals maximize earnings, we have that,

$$D = 1\left\{W_1^* > W_2^*\right\} = 1\left\{X\left(\beta_1 - \beta_2\right) - \left(\varepsilon_2 - \varepsilon_1\right) > 0\right\} \tag{18}$$

It is clear that the unobservable variable in the equation for the selection dummy, $\varepsilon_2 - \varepsilon_1$, is not independent of the unobservable in the earnings equations, $\varepsilon_1$ and $\varepsilon_2$. We have a random sample of individuals characteristics $X$ and wages $W$. Given this sample we are interested in the estimation of $\beta_1$ and $\beta_2$.

**Example 6 (Treatment effects)**: We are interested in evaluating the effect on firm capital investment of a policy that provides a certain subsidy to investment. Let $Q_1^*$ and $Q_0^*$ be a firm's amount of investment if it receives treatment (the subsidy) and if it does not, respectively. $Q_1^*$ and $Q_0^*$ are latent variables. The *Treatment Effect* (TE) for an individual firm is defined as $TE = Q_1^* - Q_0^*$. We are interested in the estimation of the *Average Treatment Effect* (ATE), that is defined as $ATE = E(Q_1^* - Q_0^*)$. We may be also interested in conditional Average Treatment Effects, $ATE(X) = E(Q_1^* - Q_0^*|X)$, where $X$ is a vector of exogenous firm characteristics. We have a random sample of firms. Each firm is observed only once, either under treatment ($D = 1$) or not ($D = 0$). That is, if $D = 1$ we observe

$Q = Q_1^*$, and if $D = 0$ we observe $Q = Q_0^*$. Typically, participation in the subsidy program is not completely random. We do not have a perfect experimental data. The treatment dummy $D$ depends on observable characteristics $Z$ and on an unobservable $u$ that may be correlated with $Q_0^*$ or/and $Q_1^*$. We want to use our sample of $\{Q, D, X, Z\}$ to estimate consistently effect of the subsidy program on investment as measured by the unconditional or the conditional average treatment effect.

**Example 7 (Friction model):** Consider a model of capital investment similar to the one in Example 3. However, now investment is not fully irreversible and it is possible to disinvest or to sell used capital. Let $K_t$ be a firm's capital stock that is productive at period $t$. Let $\Pi(K_t, K_{t-1})$ be the (intertemporal) profit function. Profits depends both on $K_t$ and $K_{t-1}$ because the existence of adjustment costs. More specifically, there is an asymmetry between the price of new capital and the price of used capital, or in other words, between the cost of capital when $K_t > K_{t-1}$, and the cost of capital when $K_t < K_{t-1}$.

$$\Pi(K_t, K_{t-1}) = \begin{cases} \Pi^{(+)}(K_t, K_{t-1}) & \text{if} \quad K_t \geq K_{t-1} \\ \\ \Pi^{(-)}(K_t, K_{t-1}) & \text{if} \quad K_t \leq K_{t-1} \end{cases} \tag{19}$$

Functions $\Pi^{(+)}$ and $\Pi^{(-)}$ are continuous, differentiable, and strictly concave in $K_t$. Profit function $\Pi$ is continuous everywhere, but it has a kink (i.e., a point of non-differentiability) at $K_t = K_{t-1}$. Define $K_t^{(+)} \equiv \arg\max_k \Pi^{(+)}(k, K_{t-1})$, and $K_t^{(-)} \equiv \arg\max_k \Pi^{(-)}(k, K_{t-1})$. Under the previous conditions, it is straightforward to show that $K_t^{(+)} < K_t^{(-)}$, and the optimal amount capital at period $t$ is:

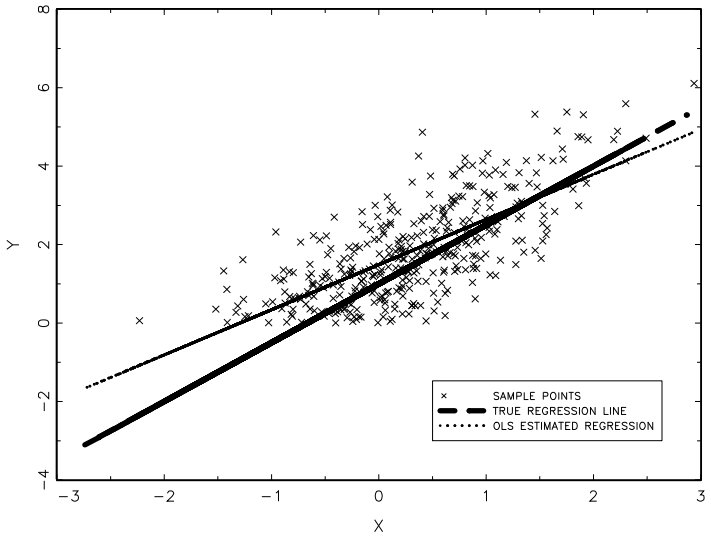
$$K_t = \begin{cases} K_t^{(+)} & \text{if} \qquad K_{t-1} < K_t^{(+)} \\ \\ K_{t-1} & \text{if} \quad K_t^{(+)} \leq K_{t-1} \leq K_t^{(-)} \\ \\ K_t^{(-)} & \text{if} \qquad K_{t-1} > K_t^{(+)} \end{cases} \tag{20}$$

The model is completed with the specification of $K_t^{(+)}$ and $K_t^{(-)}$ in terms of observable and unobservables. For instance, $K_t^{(+)} = \alpha^{(+)} + X_t\beta + \varepsilon_t$, and $K_t^{(-)} = \alpha^{(-)} + X_t\beta + \varepsilon_t$, where $\alpha^{(+)}$ and $\alpha^{(-)}$ are parameters and $\alpha^{(+)} < \alpha^{(-)}$. Given a random sample of $\{K_t, K_{t-1}, X_t\}$, we are interested in the estimation of the parameters $\alpha^{(+)}$, $\alpha^{(-)}$ and $\beta$.

# 2 Estimation of the Truncated Regression Model

## 2.1 Bias of the OLS Estimator

Consider a truncated regression model described by the expression $(Y, X) = \{(Y^*, X^*) \mid Y^* > c\}$, with $Y^* = X^*\beta + \varepsilon$. Since the constant $c$ is known, we can make $c = 0$ without lost of generality.[5] Suppose that we run an OLS regression of $Y$ on $X$. The following figure illustrates graphically the bias of the OLS estimator. The true slope of the regression line is 1.5, and the OLS estimate of this slope is 1.15 ($s.e. = 0.05$).[6]




More formally, we have that $Y = \{Y^* | Y^* > 0\} = X^*\beta + \varepsilon^{Trun}$, where $\varepsilon^{Trun} \equiv \{\varepsilon | Y^* > 0\}$. Therefore,

$$E(Y \mid X) \;=\; E(X^*\beta + \varepsilon^{Trun} \mid X) = X\ \beta + E(\varepsilon^{Trun}|X) \tag{21}$$

The term $E(\varepsilon^{Trun}|X)$ is the *sample selection term* in the conditional mean of $Y$ given $X$. Note that $E(\varepsilon^{Trun}|X) = E(\varepsilon|\varepsilon > -X\beta)$, that in general is not zero and it depends on $X$. If $\varepsilon$ is independent of $X$, the sample selection term depends on $X$ only through the index $X\beta$. Then, we can represent the selection term as a function $s(X\beta)$. It is simple to show that the

[5] If $c$ is not zero, we can always re-define $Y^*$ as the original $Y^*$ minus $c$.

[6] The DGP is such that $X^*$ and $\varepsilon$ are independent standard normal, $Y^* = 1.0 + 1.5 * X^* + \varepsilon$, and the left-truncation point is at $y = 0$. The sample size is $n = 500$.

selection term $s(X\beta)$ is a decreasing function of the index $X\beta$. To see this, note that:

$$\begin{cases} \text{As } X\beta \to +\infty, & s(X\beta) \to E(\varepsilon|\varepsilon > -\infty) = E(\varepsilon) = 0 \\ \text{As } X\beta \to -\infty, & s(X\beta) \to E(\varepsilon|\varepsilon > +\infty) = +\infty \end{cases} \tag{22}$$

Therefore, $s(X\beta)$ is negatively related with $X\beta$. In a right-truncated regression model, the selection term is also a decreasing function of $X\beta$. Taking into account that $E(Y|X) = X\beta + s(X\beta)$, we can write the following regression equation for $Y$ on $X$:

$$Y = X\ \beta + s(X\beta) + \tilde{\varepsilon} \tag{23}$$

The error term of this regression, $\tilde{\varepsilon}$, is equal to $\varepsilon^{Trun} - s(X\beta)$, and by construction it is mean independent of $X$. This expression shows the inconsistency of an OLS regression that ignores sample selection. Ignoring sample selection implies that the error term in the regression is $s(X\beta) + \tilde{\varepsilon}$, and this error term is negatively correlated with $X\beta$.

## 2.2 Maximum Likelihood Estimation

To obtain a MLE of $\beta$ we have to incorporate an additional assumption into the model: a parametric assumption on the distribution of $\varepsilon$. The typical assumption in this class of models is that $\varepsilon$ is i.i.d. over observation with a distribution $N(0, \sigma^2)$. Then, the log-likelihood function of this model and data is: $l(\beta, \sigma) = \sum_{i=1}^{n} \ln \Pr(Y = y_i \mid X = x_i)$, where the conditional probabilities have the following form:

$$\begin{aligned} \Pr(Y = y_i|X = x_i) &= \Pr(Y^* = y_i \mid X = x_i\ ;\ Y^* > 0) \\ &= \frac{\Pr(\varepsilon = y_i - x_i\beta)}{\Pr(\varepsilon > -x_i\beta)} \\ &= \frac{\frac{1}{\sigma}\phi\left(\frac{y_i - x_i\beta}{\sigma}\right)}{\Phi\left(\frac{x_i\beta}{\sigma}\right)} \end{aligned} \tag{24}$$

where $\phi(.)$ and $\Phi(.)$ are the PDF and the CDF of the standard normal. Therefore, the log-likelihood can be written as follows,

$$l(\beta, \sigma) = -n\ \ln(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i\beta)^2 - \sum_{i=1}^{n}\ln\Phi\left(\frac{x_i\beta}{\sigma}\right) \tag{25}$$

The first term in this function is the log-likelihood of the classical linear regression model. The second term accounts for truncation. Note that, in contrast to the case of a binary choice model, the log-likelihood not only depends on $\beta/\sigma$, but it depends separately on $\beta$ and $\sigma$ such that both can be identified.[7]

The log-likelihood $l(\beta, \sigma)$ is not globally concave in $(\beta, \sigma)$. This is an important issue. The maximizing of globally concave functions is a very simple task, i.e., we can use simple algorithms such as Newton, or BHHH. However, the maximization of non-globally concave functions is computationally more complicated because it requires global search over the parameter space in order to guarantee that our estimate is really the global maximum and not just a local maximum. However, for this model, it is simple to re-parameterized the log-likelihood to get a globally concave function. Define the parameters $\theta = 1/\sigma$, and $\gamma = \beta/\sigma$, and consider the log-likelihood in terms of these parameters:

$$l(\gamma, \theta) \quad = \quad n \ \ln(\theta) - \frac{1}{2} \sum_{i=1}^{n} (\theta \ y_i - x_i \gamma)^2 - \sum_{i=1}^{n} \ln \Phi (x_i \gamma)$$

The function $l(\gamma, \theta)$ is globally concave in $(\gamma, \theta)$. Note that there is a one-to-one relationship between $(\gamma, \theta)$ and $(\beta, \sigma)$. Therefore, by the *invariance-to-reparameterization* property of maximum likelihood estimation, the MLE of $(\beta, \sigma)$ is $\hat{\sigma}_{MLE} = 1/\hat{\theta}$ and $\hat{\beta}_{MLE} = \hat{\gamma}/\hat{\theta}$. The variance matrix can be obtained using the delta method.

In the context of linear regression models, the OLS estimator is consistent as long as the regressors are not correlated with the error term. Consistency of the OLS estimator is robust to heterocedasticity, serial correlation, and non-normality of the error term. Heterocedasticity is a very common feature in most cross-sectional data sets. Therefore, a relevant question is whether the previous MLE is robust to heterocedasticity in $\varepsilon$. Is this estimator still consistent when $\varepsilon$ is heterocedastic but the likelihood function is the one of an homocedastic model? The answer is no. In fact, several Monte Carlo studies have shown that the estimator can be seriously biased. This issue motivates the study of other estimators which are robust to heterocedasticity and non-normality of the disturbance.

[7] Note also that in this model we can obtain residuals $\hat{\varepsilon}$ which are consistent estimates of the errors $\varepsilon$.

## 2.3 Symmetrically Trimmed Least Squares

James Powell's work was seminal for the semiparametric estimation of truncated and censored regression models. Powell (1984) proposes Least Absolute Deviations (LAD) estimators which are robust to heterocedasticity and non-normality. Powell (1986) proposes other robust estimator based upon symmetric truncation (or censoring) of the tails of the distribution of the dependent variable. Here I describe this Symmetrically Trimmed Least Squares (STLS) estimator.

Consider a left-truncated regression model and define the following dependent variable, [8]

$$\begin{aligned} \tilde{Y} &\equiv \{Y^* \mid 0 < Y^* < 2X\beta\} \\ &= \{Y \mid Y < 2X\beta\} \end{aligned} \tag{26}$$

The variable $\tilde{Y}$ is truncated to the left and to the right. Note that the truncation points of $\tilde{Y}$ (i.e., 0 and $2X\beta$) are equidistant to the conditional mean $E(Y^*|X) = X\beta$. Given this "symmetric trimming", we have that:

$$\begin{aligned} E\left(\tilde{Y} \mid X\right) &= E\left(X\beta + \varepsilon \mid X,\ 0 < X\beta + \varepsilon < 2X\beta\right)) \\ &= X\beta + E\left(\varepsilon \mid X,\ -X\beta < \varepsilon < X\beta\right) \end{aligned} \tag{27}$$

In a linear regression of $\tilde{Y}$ on $X$, the term $E(\varepsilon \mid X, -X\beta < \varepsilon < X\beta)$ represents the *sample selection term*. It should be clear that this selection term is zero if the density of $\varepsilon$ is symmetrically distributed around zero.

Therefore, we could obtain a consistent estimator of $\beta$ by running an OLS regression of $\tilde{Y}$ on $X$. This estimator is robust to heterocedastic in $\varepsilon$. Furthermore, the symmetry assumption on the distribution of $\varepsilon$ is more general than the normality assumption. However, we do not observe $\tilde{Y}$. In order to obtain a random sample of $\tilde{Y}$ we have to truncate the observed dependent variable $Y$ to the right at $2X\beta$. But $\beta$ is unknown. To deal with this issue, we can consider the following sample criterion function:

$$Q(\beta) = \sum_{i=1}^{n} 1\left\{y_i < 2x_i\beta\right\}\left(y_i - x_i\beta\right)^2 \tag{28}$$

[8] Similarly, for a right-truncated regression-model, we define $\tilde{Y} \equiv \{Y^* \mid 2X\beta < Y^* < 0\} = \{Y \mid 2X\beta < Y\}$.

This function is the symmetrically-trimmed residual sum of squares. The STLS estimator is defined as the value of $\beta$ that minimizes this criterion. The estimator is consistent an asymptotically normal. The asymptotic variance matrix of the STLS estimator is:

$$\begin{array}{rcl} V(\hat{\beta}_{STLS}) & = & = C^{-1} D\ C^{-1} \\ \text{where:} & & \\ C & = & E\left(1\{Y < 2X\beta\}\ XX'\right) \\ D & = & E\left(1\{X\beta > 0\}\ \min\{\varepsilon^2; (X\beta)^2\}\ XX'\right) \end{array} \tag{29}$$

Note that the function $Q(\beta)$ is discontinuous and non-differentiable with respect to $\beta$ at many different points (as many as sample points). Therefore, the minimization of this criterion function may be complicated. A simple method to compute a (local) minimum is the following. Step 1: start with an initial value of $\beta$, say $\hat{\beta}^{(1)}$. For instance, the OLS estimator when we use the whole sample of $\{y_i, x_i\}$. Step 2: obtain the trimmed variable $\tilde{y}_i^{(1)} = \{y_i | y_i < 2x_i\hat{\beta}^{(1)}\}$. That is, we eliminate all the observations with $y_i > 2x_i\hat{\beta}^{(1)}$. Step 3: run an OLS regression of $\tilde{y}_i^{(1)}$ on $x_i$ to obtain a new value of $\beta$, $\hat{\beta}^{(2)}$. Iterate in Steps 2 and 3 until convergence, i.e., until $||\hat{\beta}^{(k)} - \hat{\beta}^{(k-1)}|| < small\ value$. Upon convergence, this procedure provides a local minimum of $Q(\beta)$. To check for global minimization, we have to implement a global search by applying this procedure with different initial values of $\beta$.

This method is straightforward and particularly useful when we have a large sample and the magnitude of truncation is not too severe. For relatively small samples or with severe amount of truncation, the loss of efficiency associated with the symmetric trimming may be very important, and the estimates imprecise.

*Hausman test of heterocedasticity and non-normality.* To implement a Hausman test we need an estimator that is efficient under the $H_0$ and inconsistent under $H_1$, and a estimator that is consistent both under $H_0$ and under $H_1$. Therefore, we can use the MLE and Powell's estimator to construct a test of heterocedasticity and non-normality. The null hypothesis is $\varepsilon_i \sim iid\ N(0, \sigma^2)$, and the test statistic is:

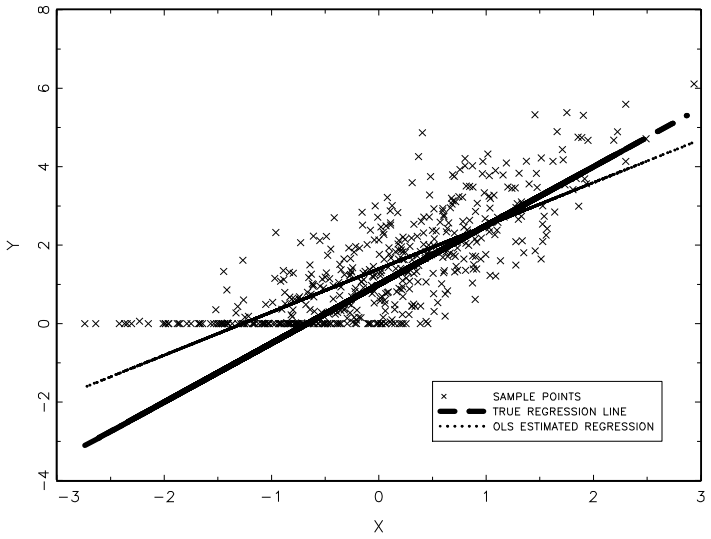
$$Hausman = (\hat{\beta}_{STLS} - \hat{\beta}_{MLE})' \left[var(\hat{\beta}_{STLS}) - var(\hat{\beta}_{MLE})\right]^{-1} (\hat{\beta}_{STLS} - \hat{\beta}_{MLE}) \tag{30}$$

that under $H_0$ is distributed as a Chi-square with $k$ degrees of freedom.

# 3 Censored Regression Model (Tobit)

## 3.1 Bias of the OLS Estimator

Consider a censored regression model such that we have a random sample of $X = X^*$, and of $Y = \max\{Y^*, c\}$, where $Y^* = X\beta + \varepsilon$. Again, we can make $c = 0$ without lost of generality. Suppose that we run a regression of $Y$ on $X$. The following figure illustrates graphically the bias of the OLS estimator. The true slope of the regression line is 1.5, and the OLS estimate of this slope is 1.10 ($s.e. = 0.04$).[9]




More formally, we have that $Y = \max\{X\beta + \varepsilon, 0\}$, or in a linear regression-like form, $Y = X\beta + \varepsilon^{Cens}$, where $\varepsilon^{Cens} \equiv \max\{\varepsilon, -X\beta\}$. Therefore,

$$E(Y \mid X) \;=\; X\beta + E(\varepsilon^{Cens} \mid X) = X\ \beta + E(\max\{\varepsilon, -X\beta\}|X) \tag{31}$$

The term $E(\varepsilon^{Cens} \mid X)$ is the *sample selection term* in the conditional mean of $Y$ given $X$. Note that $E(\varepsilon^{Cens} \mid X) = E(\max\{\varepsilon, -X\beta\}|X)$, that in general is not zero and it depends on $X$. If $\varepsilon$ is independent of $X$, the sample selection term depends on $X$ only through the index $X\beta$: i.e., $E(\varepsilon^{Cens} \mid X) = s(X\beta)$, and $s(.)$ is a decreasing function. Then, taking

[9] The DGP is such that $X^*$ and $\varepsilon$ are independent standard normal, $Y^* = 1.0 + 1.5 * X^* + \varepsilon$, and the left-censoring point is at $y = 0$. The sample size is $n = 500$.

into account that $E(Y|X) = X\beta + s(X\beta)$, we can write the following regression equation: $Y = X\beta + s(X\beta) + \tilde{\varepsilon}$, where $\tilde{\varepsilon} \equiv \varepsilon^{Cens} - s(X\beta)$ and it is mean independent of $X$. An OLS regression that ignores the sample selection term $s(X\beta)$ is inconsistent.

## 3.2 Maximum Likelihood Estimation

The log-likelihood function of this model and data is: $l(\beta, \sigma) = \sum_{i=1}^{n} \ln \Pr(Y = y_i|X = x_i)$, where the conditional probabilities have the following form:

$$\Pr(Y = y_i \mid X = x_i) \;=\; \begin{cases} \Pr(Y^* = y_i \mid X = x_i) \;=\; f_\varepsilon(y_i - x_i\beta) & \text{if} \quad y_i > 0 \\ \Pr(Y^* < 0 \mid X = x_i) \;=\; F_\varepsilon(-x_i\beta) & \text{if} \quad y_i = 0 \end{cases} \tag{32}$$

Under the assumption $\varepsilon_i \sim iid\ N(0, \sigma^2)$, the log-likelihood is:

$$l(\beta, \sigma) \;=\; -n_1\ \ln(\sigma) - \frac{1}{2\sigma^2} \sum_{y_i > 0} (y_i - x_i\beta)^2 + \sum_{y_i = 0} \ln \Phi\left(\frac{-x_i\beta}{\sigma}\right) \tag{33}$$

where $n_1$ is the number of observation with $y_i > 0$. All the comments I have made about the MLE in the truncated regression model apply also in the censored model.

## 3.3 Symmetrically Trimmed Least Squares

Consider a left-censored regression model, and define the following dependent variable:

$$\tilde{Y} = \begin{cases} 0 & \text{if} \quad Y^* \leq 0 \\ Y^* & \text{if} \quad 0 < Y^* < 2X\beta \quad = \min\{Y\ ;\ 2X\beta\} \\ 2X\beta & \text{if} \quad Y^* \geq 2X\beta \end{cases} \tag{34}$$

The variable $\tilde{Y}$ is censored both to the left and to the right. It should be clear that the censoring points of $\tilde{Y}$ (i.e., 0 and $2X\beta$) are equidistant to $X\beta$, the conditional mean of $Y^*$. Given this symmetric censoring, we have that $\tilde{Y} = \min\{Y\ ;\ 2X\beta\} = \min\{\max\{X\beta + \varepsilon; 0\}; 2X\beta\}$.[10] Or in a linear-regression-like format,

$$\tilde{Y} = X\beta + 1\{\varepsilon < -X\beta\}(-X\beta) + 1\{\varepsilon > X\beta\}(X\beta) + 1\{-X\beta \leq \varepsilon \leq -X\beta\}\varepsilon \tag{35}$$

[10] Note that $\max\{X\beta + \varepsilon; 0\} = X\beta + \max\{\varepsilon; -X\beta\}$. Therefore, $\min\{\max\{X\beta + \varepsilon; 0\}; 2X\beta\} = \min\{X\beta + \max\{\varepsilon; -X\beta\}; 2X\beta\} = X\beta + \min\{\max\{\varepsilon; -X\beta\}; X\beta\}$. Or what is equivalent, $X\beta + 1\{\varepsilon < -X\beta\}(-X\beta) + 1\{\varepsilon > X\beta\}(X\beta) + 1\{-X\beta \leq \varepsilon \leq -X\beta\}\varepsilon$.

In a linear regression of $\tilde{Y}$ on $X$, the selection term is the expected value conditional on $X$ of the error term $1\{\varepsilon < -X\beta\}(-X\beta) + 1\{\varepsilon > X\beta\}(X\beta) + 1\{-X\beta \leq \varepsilon \leq -X\beta\}\varepsilon$. As in the truncated case, this selection term is zero if the density of $\varepsilon$ is symmetrically distributed around zero.

The STLS estimator of the censored regression model is defined as the value of $\beta$ that minimizes the following criterion function:

$$Q(\beta) = \sum_{i=1}^{n} \left(\min\left\{y_i \ ; \ 2x_i\beta\right\} - x_i\beta\right)^2 \tag{36}$$

This criterion function is the residual sum of squares in the linear regression of $\tilde{Y}$ on $X$. The estimator is consistent, and asymptotically normal, and it is robust to non-normality and heterocedasticity in $\varepsilon$.

# 4 Sample Selection Models

Consider a sample selection model where $Y = (1 - D)Y_0^* + DY_1^*$, where:

$$\begin{aligned} Y_0^* &= X\beta_0 + \varepsilon_0 \\ Y_1^* &= X\beta_1 + \varepsilon_1 \end{aligned} \tag{37}$$

and

$$D = 1\{Z\gamma - u > 0\} \tag{38}$$

The unobservables $\varepsilon_0$, $\varepsilon_1$, and $u$ are not independently distributed. For instance, suppose that $D$ is the indicator of the event "the individual belongs to a union", $Y_1^*$ is the wage of the individual if he is unionized, and $Y_0^*$ represents his wage when non-unionized. We are interested in the estimation of the parameters $\beta_0$ and $\beta_1$. Sometimes, we may be interested more specifically in the average treatment effect $ATE(X) = X(\beta_1 - \beta_0)$, i.e., the average return to unionization of an individual with characteristics $X$.

## 4.1 Bias of the OLS Estimator

It is possible to construct two the following OLS estimators of the vectors $\beta_0$ and $\beta_1$: (a) a *joint OLS estimator,* where we run an OLS regression of $Y$ on $X$ and $DX$, i.e., $Y =$

$X\beta_0 + DX(\beta_1 - \beta_0) + e$; (b) *separate OLS regressions*, i.e., a regression $Y = X\beta_0 + e_0$ using the subsample of observations with $D = 0$, and a regression $Y = X\beta_1 + e_1$ using the subsample of observations with $D = 1$. It should be clear that if there are not cross-equation restrictions between the parameters $\beta_0$ and $\beta_1$, the two OLS estimators are identical, and therefore we can concentrate in only one of them, say (b).

By construction, the error term $e_j$ is $e_j \equiv \{\varepsilon_j | D = j\}$. Therefore,

$$\begin{aligned} E(e_0|X) &= E(\varepsilon_0 \mid X, D = 0) = E(\varepsilon_0 \mid X,\ u \geq Z\gamma) \\ E(e_1|X) &= E(\varepsilon_1 \mid X, D = 1) = E(\varepsilon_1 \mid X,\ u < Z\gamma) \end{aligned} \tag{39}$$

If $\varepsilon's$ and $u$ are not independent, and unless $X$ and $Z$ are independent (which is extremely unrealistic with non-experimental data), these selection terms are correlated with $X$. Therefore, the error terms $e_0$ and $e_1$ are correlated with $X$, and these OLS estimators provide inconsistent estimates of $\beta_0$ and $\beta_1$.

Let us interpret this bias in the context of the example of the return to unionization. The OLS estimation of $\beta_1 - \beta_0$, in the regression $Y = X\beta_0 + DX(\beta_1 - \beta_0) + e$, is the combination of two effects: (1) the actual return to unionization, $\beta_1 - \beta_0$; and (2) the fact that those workers who decide to be unionized tend to be the ones who have larger "treatment effect" or wage differential $Y_1^* - Y_0^*$. The first factor is the causal effect that we want to estimate. The second factor is spurious, it is not a causal effect of unionization. For the sake of illustration, suppose that $X$ is just a constant term. Suppose also that unionization has two effects: it increases the constant term, i.e., $\beta_1 > \beta_0$, and it reduces wage dispersion, i.e., $\varepsilon_1 = \lambda\varepsilon_0$ where $\lambda < 1$. Also, suppose that the only factor that affects the unionization decision is the wage differential (i.e., Roy model) such that $Z\gamma - u = Y_1^* - Y_0^* = (\beta_1 - \beta_0) + (\varepsilon_1 - \varepsilon_0) = (\beta_1 - \beta_0) - (1 - \lambda)\varepsilon_0$. In this example, it is clear that:

$$\begin{aligned} p\lim\ \hat{\beta}_0^{OLS} &= E(Y|D = 0) = \beta_0 + E\left(\varepsilon_0 \mid \varepsilon_0 > \frac{\beta_1 - \beta_0}{1 - \lambda}\right) > \beta_0 \\ p\lim\ \hat{\beta}_1^{OLS} &= E(Y|D = 1) = \beta_1 + \lambda\ E\left(\varepsilon_0 \mid \varepsilon_0 < \frac{\beta_1 - \beta_0}{1 - \lambda}\right) < \beta_1 \end{aligned} \tag{40}$$

Therefore, in this example, $\hat{\beta}_0^{OLS}$ overestimates $\beta_0$ because non-unionized workers have higher

values of $\varepsilon_0$ (e.g., higher productivity), $\varepsilon_0 > \dfrac{\beta_1 - \beta_0}{1-\lambda}$. Also, $\hat{\beta}_1^{OLS}$ underestimates $\beta_1$ because unionized workers have lower values of $\varepsilon_0$, i.e., $\varepsilon_0 < \dfrac{\beta_1 - \beta_0}{1-\lambda}$. As a result, under the previous assumptions, the OLS estimator of $\beta_1 - \beta_0$ underestimates the true return of unionization.

## 4.2 Maximum Likelihood Estimation

The dependent variables of the model are $Y$ and $D$, and the exogenous explanatory variables are $X$ and $Z$. The log-likelihood function of this model and data is,

$$l(\beta, \gamma, \Omega) = \sum_{i=1}^{n} \ln \Pr(Y = y_i, D = d_i \mid X = x_i, Z = z_i) \tag{41}$$

with probabilities,

$$\begin{aligned} \Pr(Y = y_i, D = 0 \mid X = x_i, Z = z_i) &= \Pr(\varepsilon_0 = y_i - x_i\beta_0 \ ; \ u_i > z_i\gamma) \\ &= \int_{z_i\gamma}^{+\infty} f_{\varepsilon_0,u}(y_i - x_i\beta_0, u) \ du \end{aligned} \tag{42}$$

and

$$\begin{aligned} \Pr(Y = y_i, D = 1 \mid X = x_i, Z = z_i) &= \Pr(\varepsilon_1 = y_i - x_i\beta_1 \ ; \ u_i < z_i\gamma) \\ &= \int_{-\infty}^{z_i\gamma} f_{\varepsilon_1,u}(y_i - x_i\beta_1, u) \ du \end{aligned} \tag{43}$$

where $f_{\varepsilon_0,u}$ and $f_{\varepsilon_1,u}$ are the joint densities of $(\varepsilon_0, u)$ and $(\varepsilon_1, u)$, respectively.

When using a MLE of this model, researchers generally assume that $(\varepsilon_0, \varepsilon_1, u)$ have a joint normal distribution. The variance of $u$ is normalized to 1. The parameters that enter in this likelihood function are $\beta_0$, $\beta_1$, $\gamma$, the standard deviations $\sigma_0$ and $\sigma_1$, and the covariances $\sigma_{0u}$ and $\sigma_{1u}$. In general, this likelihood function is not globally concave and it can have several local maxima. Furthermore, in contrast to the truncated regression model and the censored regression model, there is not a reparameterization under which the likelihood is concave. Therefore, we should initialize our optimization algorithm with different values of the parameters, keep track of the likelihood values obtained upon convergence, and then compare these likelihood values to obtain (hopefully!) the global maximum.

## 4.3 Heckman's Two Step Method

Heckman (1976, 1979) proposed an alternative two-stage approach that provides consistent estimates of the sample selection model and that is very simple to implement. The computationally simplicity of this two-step method make it very attractive in applications. However, there is at least other important reason why Heckman's two-step method has been so popular in applications. As in the case of truncated and censored regression models, the MLE is not robust to heterocedasticity and non-normality. Although Heckman's two step approach was proposed in the context of a parametric model with normal and homocedastic disturbances, one of the most attractive features of this estimator is that it can be extended to a semiparametric context with non-normal and heterocedastic errors.

Let's consider first this estimator in the context of a fully parametric model with normal and homocedastic unobservables. First, note that:

$$\begin{aligned} E(Y \mid X, Z, D=0) &= X\beta_0 + E(\varepsilon_0 \mid X, Z, D=0) \\ &= X\beta_0 + \frac{1}{1-F_u(Z\gamma)} \int_{Z\gamma}^{+\infty} E\left(\varepsilon_0 | u\right)\ f_u(u)\ du \end{aligned} \tag{44}$$

and,

$$\begin{aligned} E(Y \mid X, Z, D=1) &= X\beta_1 + E(\varepsilon_1 \mid X, Z, D=1) \\ &= X\beta_1 + \frac{1}{F_u(Z\gamma)} \int_{-\infty}^{Z\gamma} E\left(\varepsilon_1 | u\right)\ f_u(u)\ du \end{aligned} \tag{45}$$

Under normality of $\{u, \varepsilon_0, \varepsilon_1\}$, these expressions become:

$$\begin{aligned} E(Y \mid X, Z, D=0) &= X\beta_0 + \sigma_{0u}\ \lambda(-Z\gamma) \\ E(Y \mid X, Z, D=1) &= X\beta_1 - \sigma_{1u}\ \lambda(Z\gamma) \end{aligned} \tag{46}$$

where the function $\lambda(c) \equiv \dfrac{\phi(c)}{\Phi(c)}$ is called Mill's inverse ratio or Heckman's lambda.

Based on this result Heckman proposed the following two step procedure. Step 1: estimate $\gamma$ by ML in the Probit model $D = 1\{Z\gamma - u > 0\}$. Obtain $\{z_i\hat{\gamma}\}$ for every observation in the sample, and compute estimates for the Heckman's lambdas, $\hat{\lambda}_{0i} = \phi(-z_i\hat{\gamma})/\Phi(-z_i\hat{\gamma})$, and $\hat{\lambda}_{1i} = \phi(z_i\hat{\gamma})/\Phi(z_i\hat{\gamma})$. Step 2: run an OLS regression for $Y$ on $X$ and $\hat{\lambda}_0$ using the subsample of observations with $D = 0$, and run an OLS regression for $Y$ on $X$ and $\hat{\lambda}_1$ using the

subsample of observations with $D = 1$. This procedure provides consistent estimates of $\beta_0$, $\beta_1$, $\sigma_{0u}$ and $\sigma_{1u}$. Amemiya (1985, pp. 370-371) provides an expression to correct standard errors of the parameter estimates taking into account the estimation error in the variables $\hat{\lambda}_0$ and $\hat{\lambda}_1$.

How are we controlling for selection bias in this procedure? We are controlling for selection bias by including in the regression the (estimated) selection term $\hat{\lambda}$. How can we identify separately the causal effect of $X$ on $Y$ (through $X\beta_j$) and the effect though the selection bias $\hat{\lambda}_j$? Or in other words, why $\hat{\lambda}$ and $X$ are not collinear? There are two possible reasons. First, there may be variables in $Z$ which are not in $X$ (i.e., exclusion restrictions). If that is the case, and if these variables have enough explanatory power in the Probit model, $\hat{\lambda}_j$ has sample variation that is *independent* of $X$. And second, $\hat{\lambda}$ is a non-linear function of $Z\hat{\gamma}$. Even if $Z \subseteq X$, the variable $\hat{\lambda}$ has sample variation that is *linearly independent* of $X$. The first source of identification is called *identification from exclusion restrictions*, and it does not depend on our functional form assumptions, i.e., we have identification even if the model specifies a nonparametric relationship between $Y_j^*$ and $X$. The second source of identification is called *identification through functional form* and it crucially depends on our parametric assumptions, i.e., linearity of the relationship between $Y_j^*$ and $X$, and normality of the disturbances.

The previous discussion illustrates an additional reason why we might be interested in relaxing the normality assumption. Even if we are interested in linear effects of $X$ on $Y_j^*$, we would like that the identification of these effects do not only rely on the linearity assumption and a parametric assumption on the distribution of the unobservables. We now describe an extension of Heckman's two stage procedure that allows for a general distribution of the unobservables. Consider the sample selection where the unobservables $(\varepsilon_0, \varepsilon_1, u)$ are independent of $(X, Z)$ and they have an arbitrary probability distribution with support the Lebesgue measure on the Euclidean space. In fact, we can allow for heterocedasticity in $(\varepsilon_0, \varepsilon_1, u)$ as long as the variances and covariances of these variables depend on $(X, Z)$ only

through the index $Z\gamma$. Without further assumptions, the model implies that:

$$\begin{aligned} E(Y \mid X, Z, D=0) &= X\beta_0 + s_0(Z\gamma) \\ E(Y \mid X, Z, D=1) &= X\beta_1 + s_1(Z\gamma) \end{aligned} \tag{47}$$

where now the functional form of the selection functions $s_0(.)$ and $s_1(.)$ is unknown. However, we know that they are single-index functions. These functions only depend on $Z\gamma$. Given an estimate of $\gamma$ from the binary choice model,[11] we can approximate arbitrarily well the terms $s_j(Z\hat{\gamma})$ using a polynomial of order $q$ in $Z\hat{\gamma}$. That is, in the second stage we can estimate by OLS the regressions:

$$\begin{aligned} \{Y|D=0\} &= X\beta_0 + \sum_{j=1}^{q} \rho_{0j} \ (Z\hat{\gamma})^j + e_0 \\ \{Y|D=1\} &= X\beta_1 + \sum_{j=1}^{q} \rho_{1j} \ (Z\hat{\gamma})^j + e_1 \end{aligned} \tag{48}$$

Some authors have proposed also two use a polynomial in estimated Heckman's lambda, or a polynomial in the estimated discrete choice probability (i.e., propensity score). We can also use other type of semiparametric estimators for partially linear models (see Robinson, 1983, and Yatchew, 2003). It is clear that the identification of $\beta_0$ and $\beta_1$ is based only on exclusion restrictions. This makes also clear that in order to identify these parameters using this approach, the index $Z\hat{\gamma}$ should have enough sample variability independent of $X$. Also, we will have to justify our exclusion restrictions based on economic arguments and our knowledge of the problem.

[11] The parametric specification of the discrete choice model is not important here. We can also use a polynomial in $z_i$ in the probit model.

# References


[1] Amemiya, T. (1985): "Advanced Econometrics," Hardvard University Press. Cambridge, Massachusetts.

[2] Heckman, J. (1976): "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Model," Annals of Economic and Social Measurement, 15, 475-492.

[3] Heckman, J. (1979): "Sample Selection Bias as a Specification Error," Econometrica, 47, 153-161.

[4] Heckman, J., and B. Honore (1990): "The Empirical Content of the Roy Model," Econometrica, 58, 1121-1149.

[5] Powell, J. (1984): "Least Absolute Deviations Estimation for the Censored Regression Model," Journal of Econometrics, 25, 303-325.

[6] Powell, J. (1986): "Symmetrically Trimmed Least Squares Estimation for Tobit models," Econometrica, 54, 1435-1460.

[7] Roy, A. D. (1951): "Some Thoughts on the Distribution of Earnings," Oxford Economic Papers (New Series), 3, 135-146.

[8] Robinson, P. (1988): "Root-N-Consistent Semiparametric Regression," Econometrica, 56, 931-54.

[9] Yatchew, A. (2003): "Semiparametric Regression for the Applied Econometrician," in Themes in Modern Econometrics, ed. P.C.B. Phillips, Cambridge University Press.